2017

Data Science for Social Good Europe

Summer Fellowship

A critical introduction to statistics
Part 2: Inference and prediction, Stats vs. ML

Momin M. Malik

v2.1, 15 August 2017

Any views expressed here are my own, and do not necessarily reflect those of DSSG.

Review from last time:

- *"briefly, and in its most concrete form, the object of statistical methods is the **reduction of data**."* –Fisher, 1922
- Statistics is the use of probability as a model for variability in the world

Learning Goals

- Know what “estimation” means in statistics, and why it is central (everything is an estimator, from the standard errors we use to do inference on other estimators, to cross-validation)
- Understand that the Central Limit Theorem applies to *distributions of* (summaries of) *distributions*
- See that inference and prediction are distinct, and that the most predictive features may not be statistically significant and, conversely, that statistical significance should not be used for feature selection towards prediction
- See that the bias-variance tradeoff means that the models that predict the best are not necessarily the most “true”

Review: Likelihood principle

- Last time, I discussed linking *data* to *probability distributions* with the *likelihood principle*.
- We have some probability distribution,

$$p(x) \propto e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- and some data, say, x_1 = -1.38, x_2 = -0.44, x_3 = 1.64, x_4 = -0.25, etc.
- The joint probability of independent events is the product,

$$p(x_1, x_2, x_3 \ldots) = p(x_1) \times p(x_2) \times p(x_3) \times \cdots$$

$$\propto e^{-\frac{(x_1-\mu)^2}{2\sigma^2}} \times e^{-\frac{(x_2-\mu)^2}{2\sigma^2}} \times e^{-\frac{(x_3-\mu)^2}{2\sigma^2}} \times \cdots$$

- Likelihood: plug in data for x_i's, treat result as function of μ and σ^2:

$$\mathcal{L}(\mu, \sigma^2) = e^{-\frac{(-1.38-\mu)^2}{2\sigma^2}} \times e^{-\frac{(-0.44-\mu)^2}{2\sigma^2}} \times e^{-\frac{(1.64-\mu)^2}{2\sigma^2}} \times \cdots$$

Review: Likelihood principle

- In practice, we use the *log likelihood*, since it has the same maximum:

$$\ell(\mu, \sigma^2) = -\frac{(-1.38-\mu)^2}{2\sigma^2} - \frac{(-0.44-\mu)^2}{2\sigma^2} - \frac{(1.64-\mu)^2}{2\sigma^2} - \dots$$

- Exercise: use calculus to maximize this (or minimize the negative)
- The solution for μ is $n^{-1}(1.38 + 0.44 + 1.64 + \dots)$, the sample mean!
- The solution for σ^2 is $\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2$.

Estimation

- What we did is actually a very narrow task: given data, find the parameters of the underlying distribution (assuming the distribution is known)
- This is known as *estimation*
- “Maximum likelihood estimation” is the dominant approach, but there are others
- (There are proofs about estimators getting you close to the true parameters)

Estimation

- How does this relate to practice?
- In stats, we imagine that there is:
  - Some underlying “true” relationship between variables (regression)
  - Some underlying “true” decision boundary (classification)
  - Some underlying “true” distribution for variables (density estimation)
- We want to use data to recover this underlying “truth”

Estimation

- Models are definitions of *estimators* of the desired object (algorithms calculate the estimator: “convex optimization” is about finding the fastest way to calculate estimators)
- E.g., *k*-nearest neighbors, SVMs, Naïve Bayes, logistic regression, random forests:
  - All these classifiers are *estimators* of the (hypothesized true underlying) decision boundary
- Different ones are optimal, and reach better results more quickly (“efficiency”), under differing scenarios

Estimation

- Why do we need multiple estimators?
- The “best” estimator depends on the phenomenon, there is no perfect general-purpose estimator
- There is also no comprehensive list of estimators
- We usually don’t know about the phenomenon in advance
- Different estimators make different assumptions that are appropriate for different phenomena

Estimation, inference, and prediction

Estimation is only one of three major tasks of statistics. The others are *inference* and *prediction*.

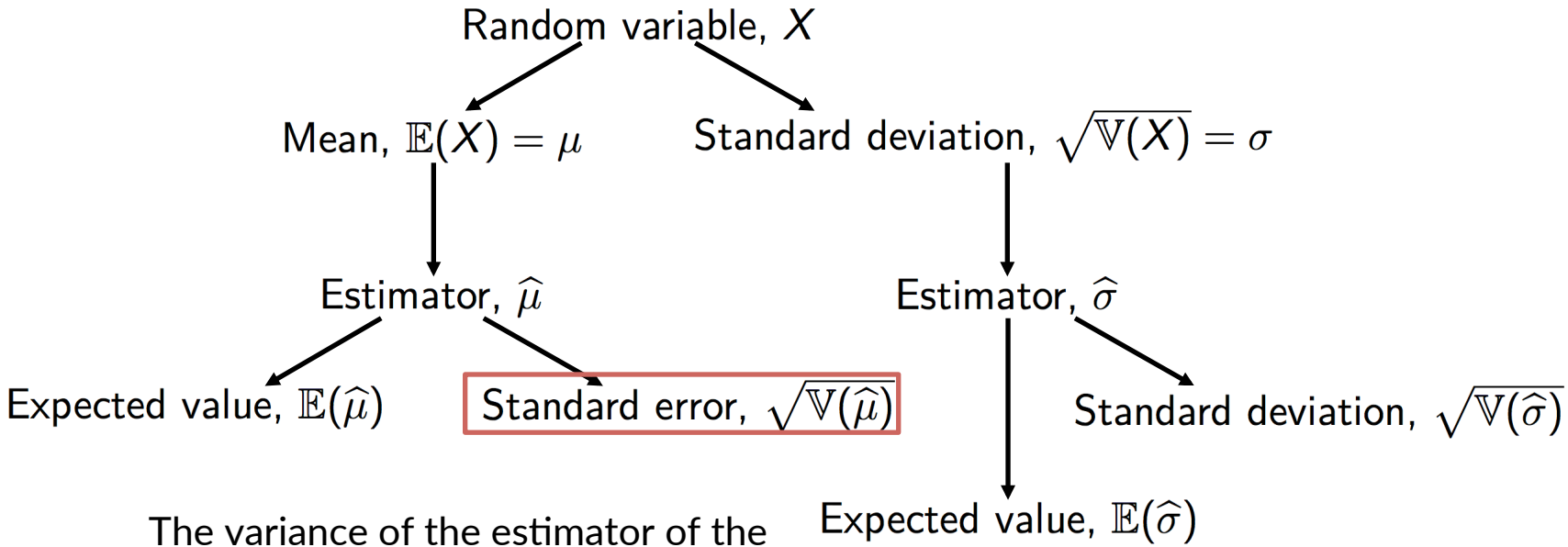
Estimation, inference, and prediction

- Inference: the other part of the quote from D. R. Cox:
  - *"Probability is used in two distinct, although interrelated, ways in statistics, phenomenologically to describe haphazard variability arising in the real world and epistemologically to represent uncertainty of knowledge."* —D. R. Cox, "Role of models in statistical analysis" (1990)
- Inference is about saying: we can always find an estimate. But is it any good?

Inference

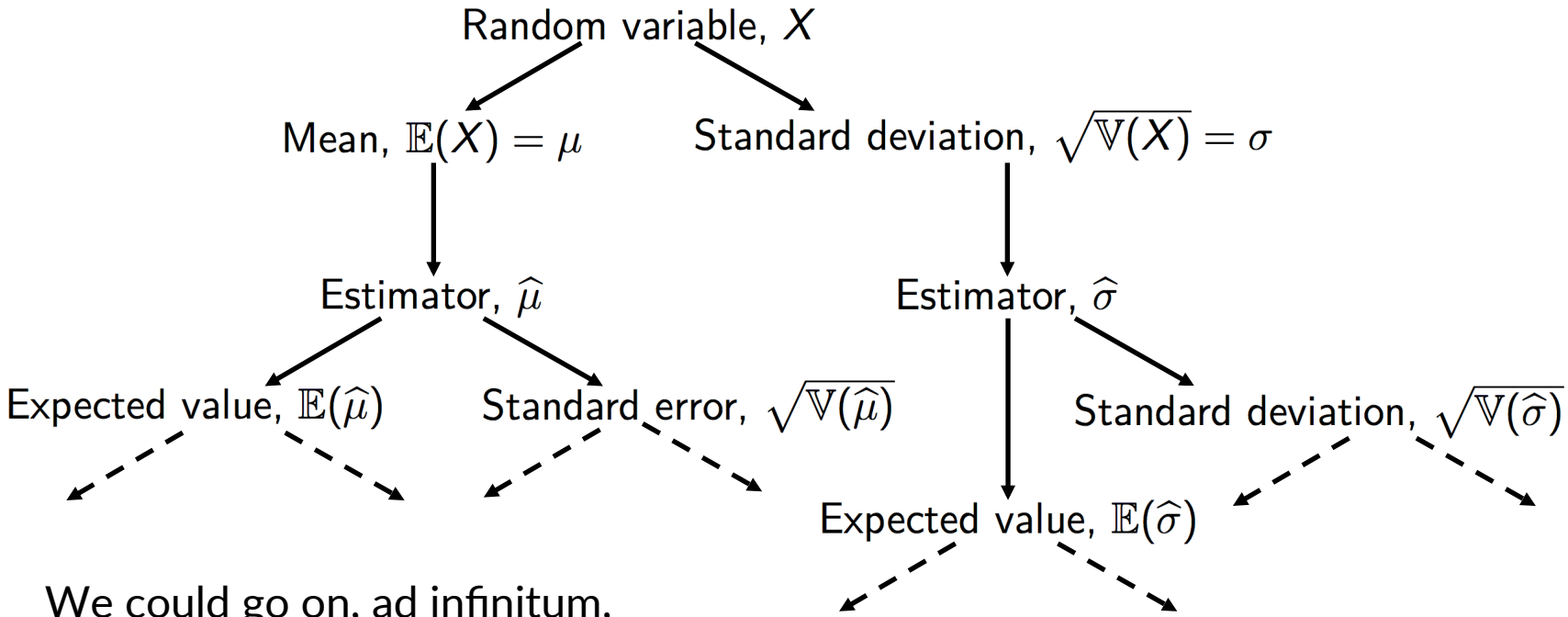
- How do we know if an estimator is any good?
- *An estimator is itself a random variable*
- That means it has a distribution!
- Studying this distribution can let us know about the performance of the estimator

Inference



The variance of the estimator of the mean gives us the uncertainty of the estimate, and is given the special name of the *standard error*. If the uncertainty is small enough, we make *inferences*.

Inference



We could go on, ad infinitum,
but we generally stop here.

Inference

- Standard errors are central to inference
- Standard errors are *about* estimators, but *themselves also have estimators*
- Inference: using one estimator to quantify the uncertainty of another estimator

Aside: Law of Large Numbers and Central Limit Theorem

Law of Large Numbers / Central Limit Theorem

- The law of large numbers says that as you get more data from the underlying phenomenon, the closer the sample mean is to the true mean
- The central limit theorem says (among other things) that not only is the sample mean close to the true mean, it is *distributed normally* around the true mean
  - Key point: the "distribution" is over *multiple data sets*
  - We seldom have multiple distinct data sets (and if we did, why wouldn't we just combine them??)
  - The set of other data sets we *could have drawn* is a theoretical construct we appeal to in frequentist inference

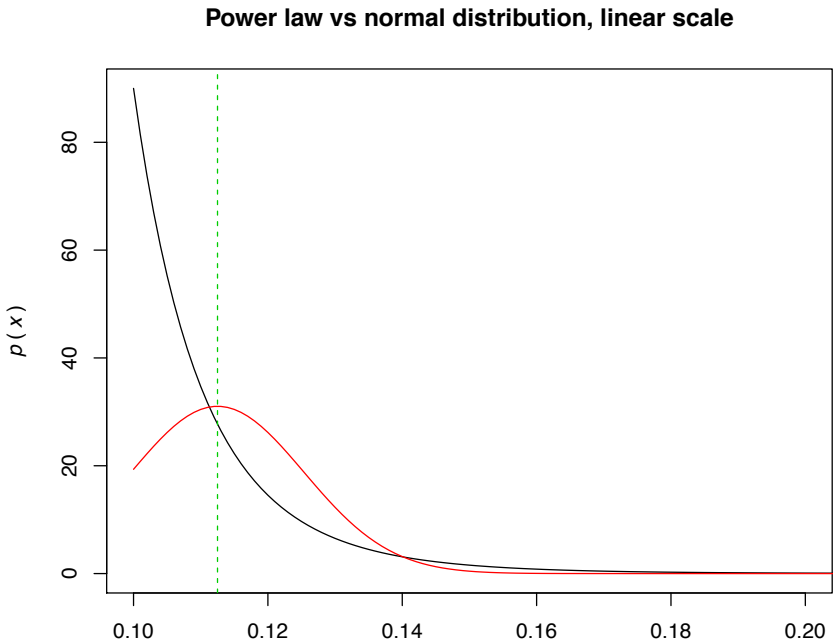
Law of Large Numbers / Central Limit Theorem

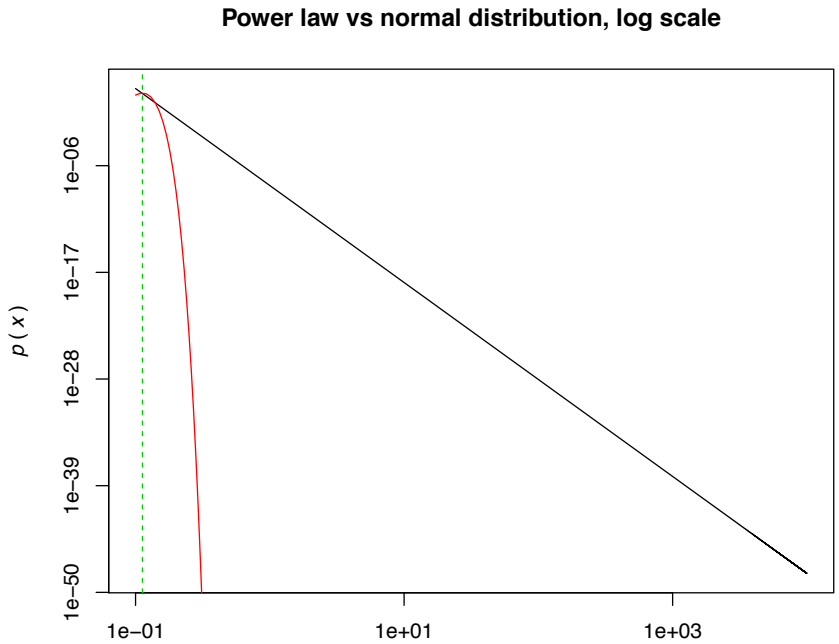
- To illustrate: take one the most “non-normally” distributed distributions, a power law distribution
- (Caution: *heavy-tailed distributions* are not automatically “power laws”!)
- For certain ranges of its parameters (x_{min} and the exponent of the power law, α), it has a mean: it’s just that the mean is not informative (it conflates x_{min} and α)
- The sample mean, over multiple samples, is still normally distributed around the true mean!

(A note on simulations)

- Simulations (producing “synthetic data”/“toy data”) are very useful in statistics
- It is to investigate, “if the world works the way I say it does, do my claims hold?” Low bar, but useful
- Important to note: Usually, we have no idea if we were successful from methods themselves
  - (In Prof. João Paulo Costeira’s lecture, he had an example where their model seemingly incorrectly said one wall of a building [Rem Koolhaas’ Casa da Musica in Porto] was not flat, but consulting blueprints showed that the model was more correct that human vision! This is the dream of modeling: we have some way of validating model output *independent of the data that goes into the model*, and that such a validation shows that our model magically beats the manual/human way of doing things. But we almost never have even the ability to do such validation—and even more rarely does it work out as such.)
- When there is no way to independently validate a model, we can *simulate the world working a specific way*
- **For frequentist claims especially, we can simulate long-run frequencies**
- (This is distinct from simulation *modeling*, which is about modeling the world rather than exploring the performance of statistical techniques)

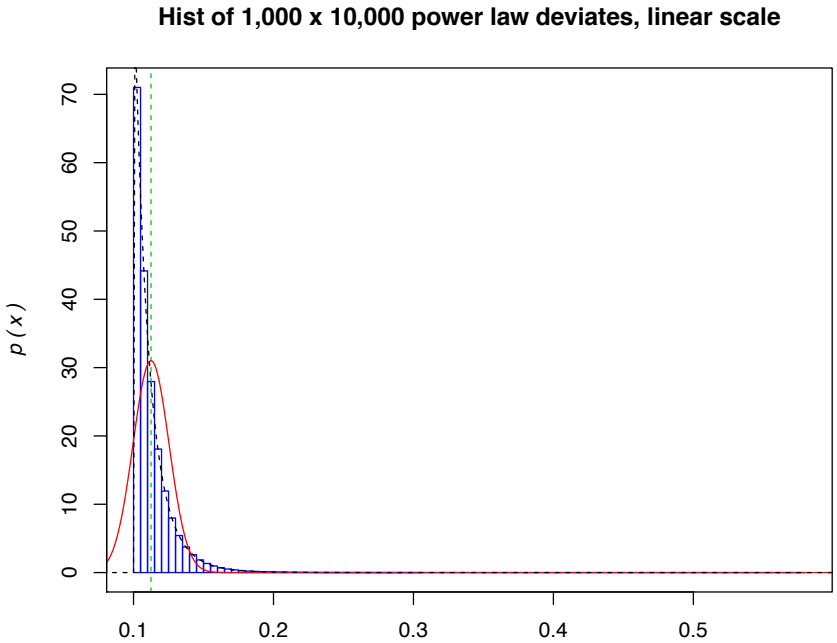
Theoretical distributions: Power law vs. normal

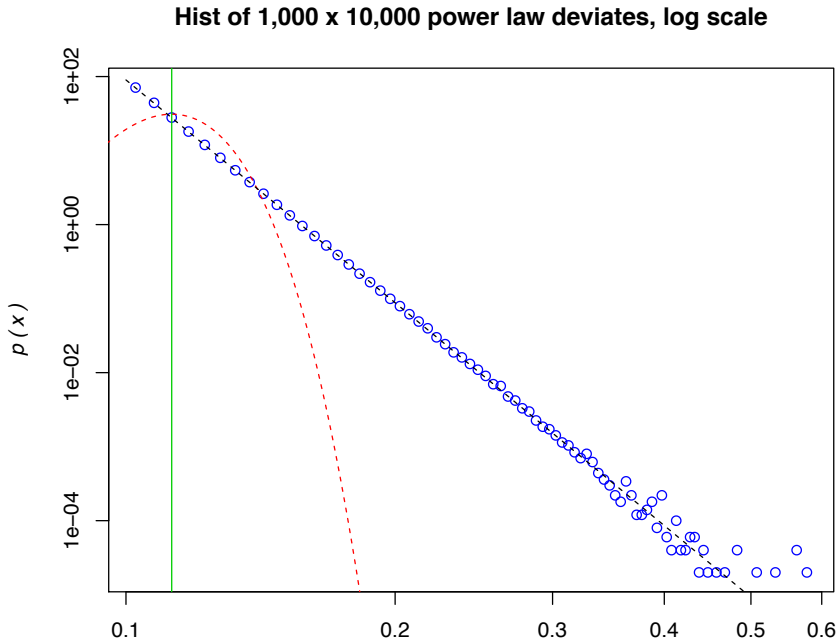




The dotted green line is the mean of the power law, the red curve is a normal distribution with the same mean and standard deviation as the power law. As we can see, the "tail" of the normal distribution drops off to being (effectively) zero rather quickly, unlike the power law distribution, which has a tail that lingers with some mass forever.

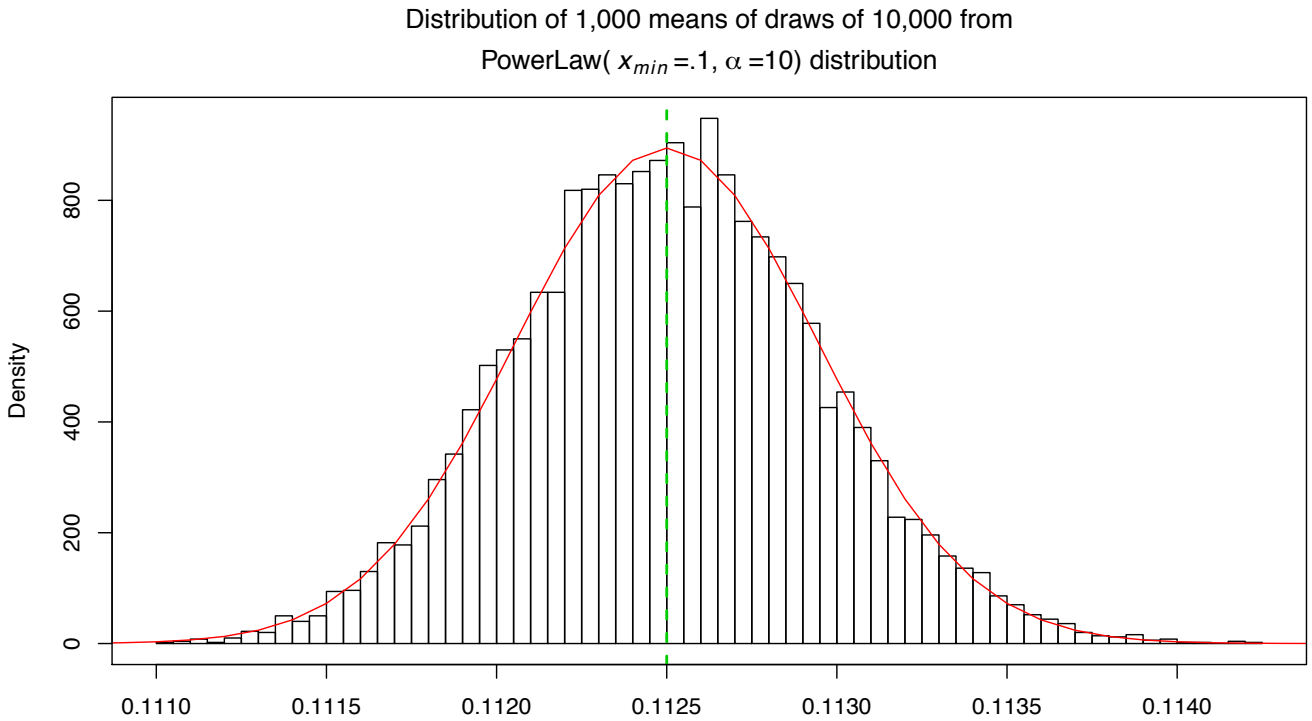
Distribution of *draws* (with normal reference)





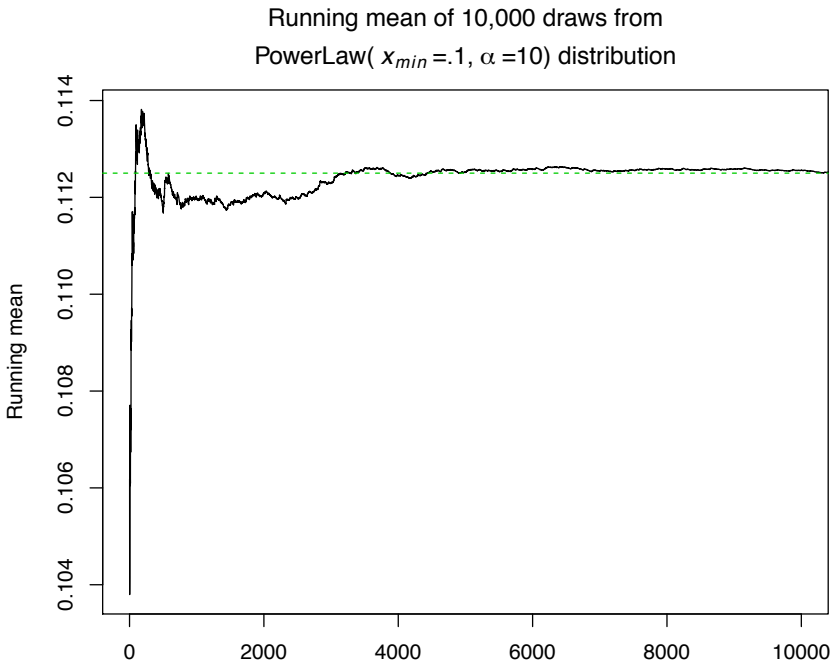
As we can see, a normal distribution is a lousy approximation for data drawn from a power law distribution, but the power law distribution that generated the data is (unsurprisingly) a very decent fit. (Also, another caution: power law ⇒ straight line in log-log scale, but straight line in log-log scale ⇏ power law. See Cosma Shalizi, “So, You Think You Have a Power Law, Do You? Well Isn't That Special,” 2010).

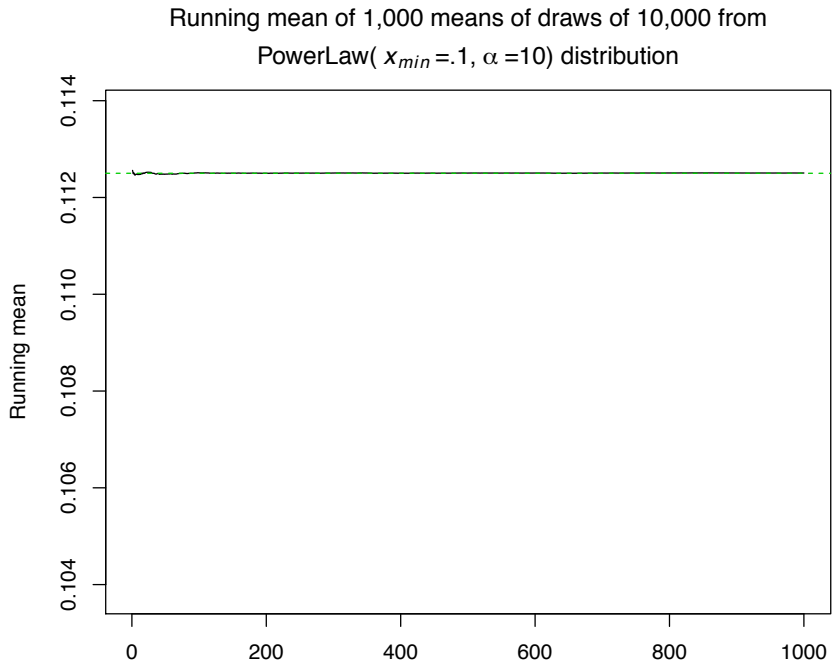
Distribution of *means* (grouped draws)



But the *means* are normally distributed! This is why we put normal-based confidence intervals (point ± 1.96*standard_error) on estimates, regardless of underlying distribution: we imagine our specific data set is one of many possible draws from an underlying process. I we put a normal-based 95% confidence interval around it, that interval will contain the true value from 19 out of 20 (95%) data sets we might see.

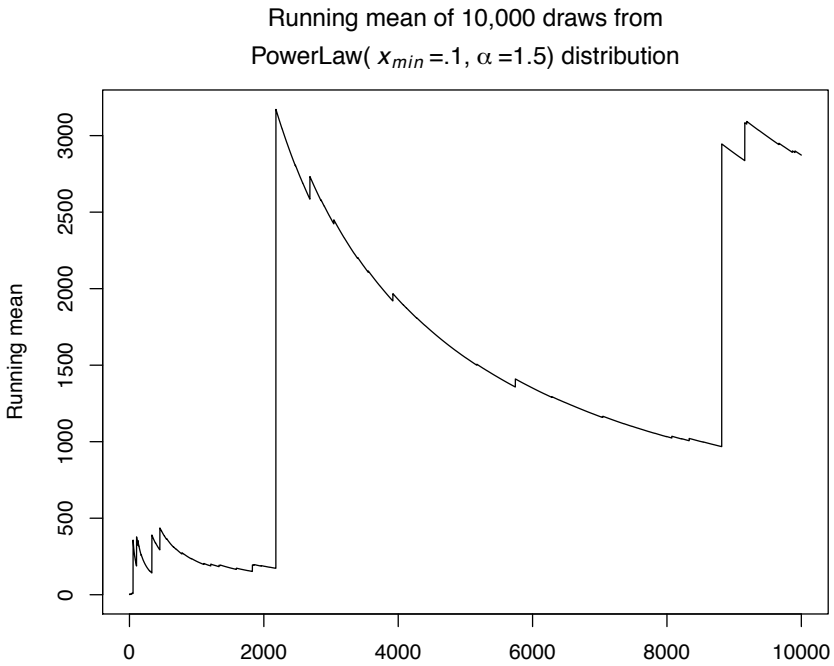
Alternative view: Mean as $n \to \infty$

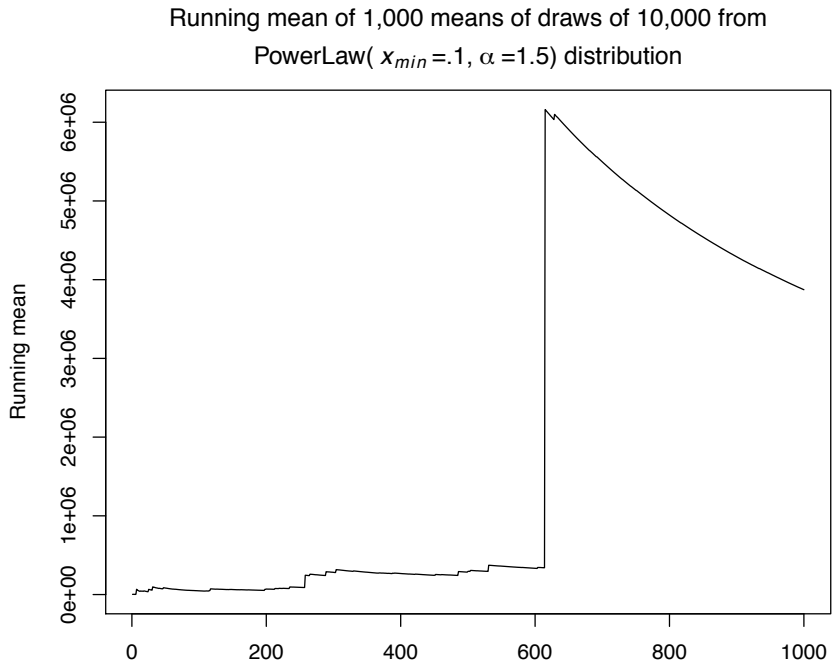




The running mean, as we get more data, approaches the underlying mean (this is the law of large numbers), shown on the left. If we take the same 10,000,000 independent draws, and instead group them and take the mean of each 10,000, we will both get to the underlying mean more quickly, shown on the right, *and* means will be normally distributed (previous slide is a histogram of the right plot's *y* values).

But the mean may not exist, in which case, no CLT





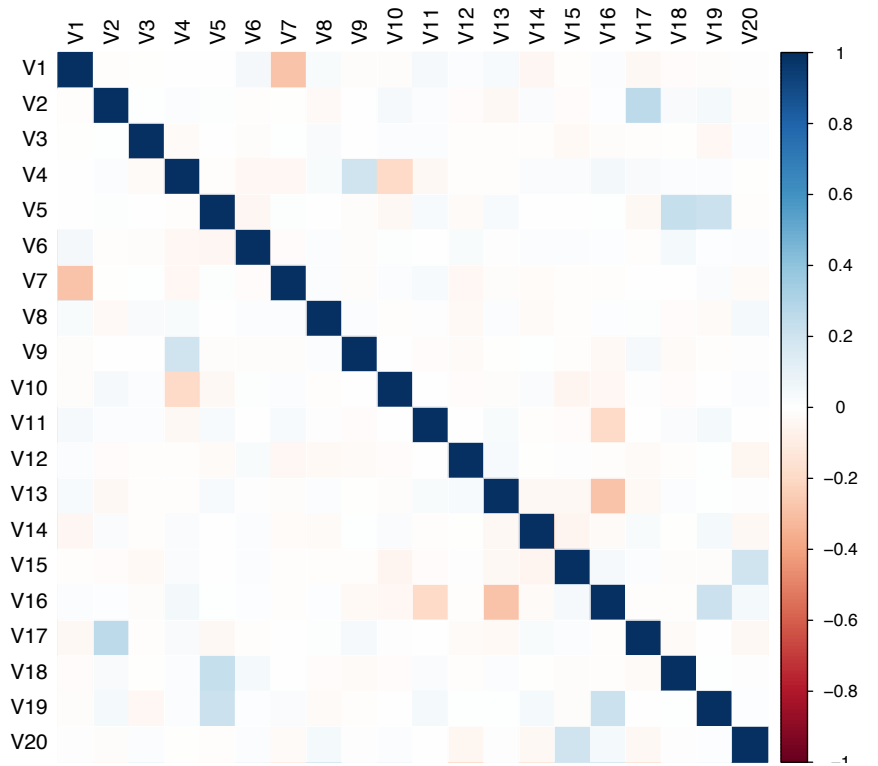
The mean of power law distributions with certain exponents is not well defined. In simulations, we see that this is because we keep seeing single observations so extreme that they alone pull the average up enormously—both the running average (left) or the mean of the means (right) of the same 10,000,000 draws.

Lessons

- If the mean exists: the sample average is normally distributed around the true mean (over multiple data sets), regardless of the underlying distribution
  - And also regardless of whether the mean is *useful* or not: for power laws (where the mean exists), it is not useful
  - (A less exotic example of the mean not being useful, or at least being misleading: bimodal distributions)
- (End aside)

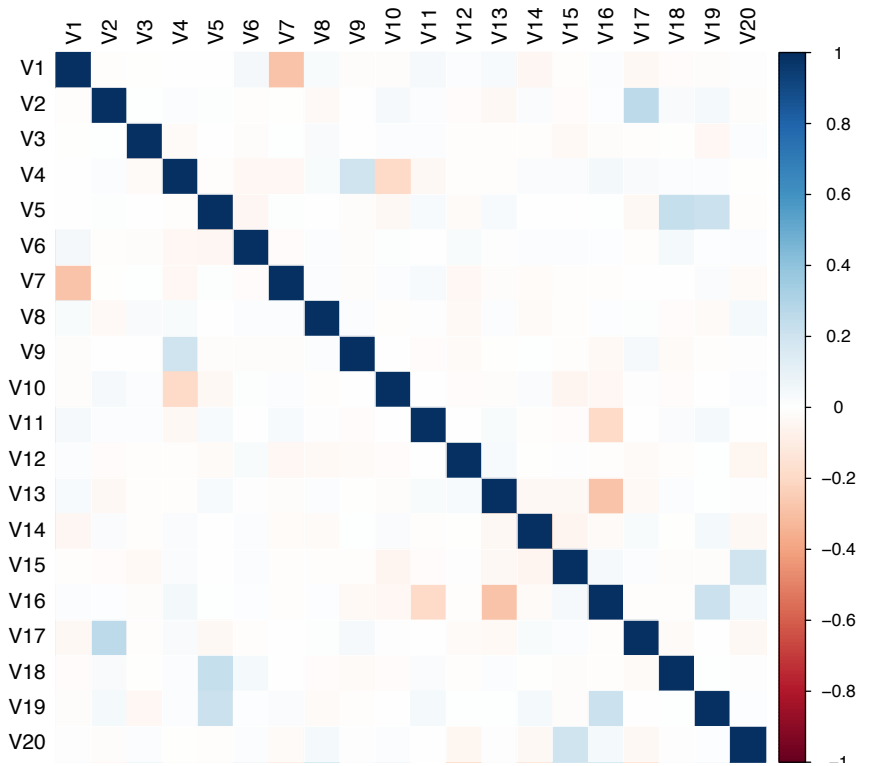
p-hacking/fishing

- Say you have 20 variables
- 20 choose 2 gives 190 pairwise interactions
- Say the 20 were *all* actually independent
- Using a test with a false positive rate of 5%, you would get ~10 “significant” bivariate relationships
- Right: I generate 20 standard normal variables independently, and look at correlations between them



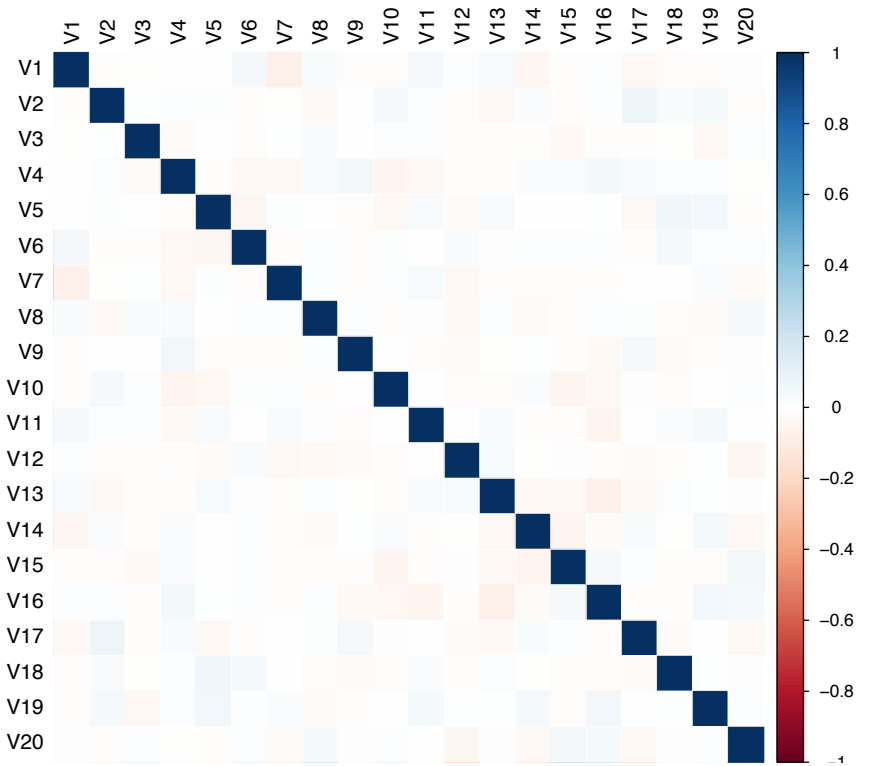
p-hacking/fishing

- This is the “multiple comparison problem,” a frequent critique of data mining
  - With enough relationships, you will find patterns
  - And with enough relationships across enough studies, plus publication bias…
  - Should put in extra controls: Bonferroni correction is simplest, divide desired level by the number of comparisons (e.g., .05/190), but very restrictive



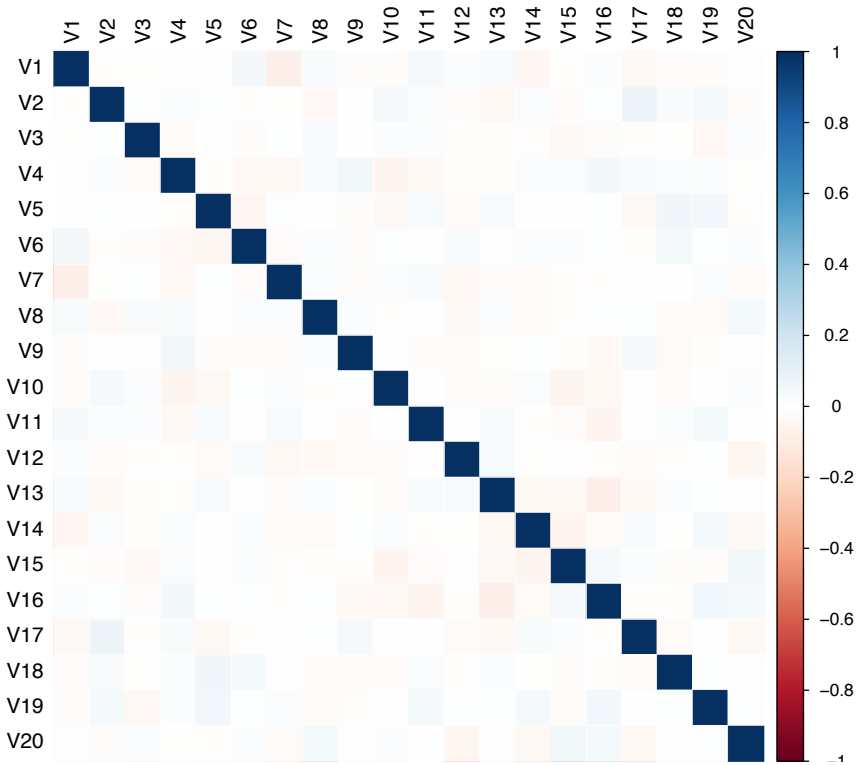
p-hacking/fishing

- This is the “multiple comparison problem,” a frequent critique of data mining
  - With enough relationships, you will find patterns
  - And with enough relationships across enough studies, plus publication bias…
  - Should put in extra controls: Bonferroni correction is simplest, divide desired level by the number of comparisons (e.g., .05/190), but very restrictive



p-hacking/fishing

- (Reminder: this is another example of a simulation to *explore the behavior of our techniques*, if the world works the way we propose it does. In reality, we could never know which inferences were true and which were spurious; we can only rely on theoretical frequentist guarantees).



Prediction

Prediction

- When we fit models, we get weights/coefficients
- So we can plug in new data into the same model
- We call what we get out “predictions”
- (I’d love somebody to look, historically and philosophically, at the emergence of prediction as a distinct task!)

Prediction is not what you think

- *"**It's not prediction at all!** I have not found a single paper predicting a **future result**. All of them claim that a prediction **could have been** made; i.e. they are post-hoc analysis and, needless to say, negative results are rare to find."*

  Gayo-Avello, "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper" (2012)

Prediction

- “Predicted values” is a technical term synonymous with “fitted values,” so in some sense Gayo-Avello is being unfair
- But when the public reads press releases about scientists successfully “predicting X,” they don’t know that
- Read “We can predict X” instead as “We found a model that fits well”
- Fitting well is still an accomplishment, but it’s quite different from actually being able to tell the future

Prediction and inference: What's the difference?

≈ Statistics and machine learning:
What's the difference?

Data Mining and Statistics: What's the Connection?

Jerome H. Friedman
Department of Statistics and
Stanford Linear Accelerator Center
Stanford University
Stanford, CA 94305
jhf@stat.stanford.edu

What is machine learning?

- *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P* ***if its performance*** *at tasks in T, as measured by P,* ***improves with experience*** *E."*

  –Tom Mitchell, 1997

- This definition has nothing to do with statistics or probability!

What is machine learning?

"Statistics is the science of learning from data. Machine learning (ML) is the science of learning from data. These fields are identical in intent although they differ in their history, conventions, emphasis and culture."

"At first, ML researchers developed expert systems that eschewed probability. But very quickly they adopted advanced statistical concepts like empirical process theory and concentration of measure. This transition happened in a matter of a few years."

–Larry Wasserman, "Rise of the Machines" (2014)

What is machine learning?

- The “learning” is a *metaphor*. The way in which machines “improve with data” has only a fleeting resemblance to human learning.
- “A.I. systems tend to be passive vessels, dredging through data in search of statistical correlations; humans are active engines for discovering how things work.” –Gary Marcus, “Artificial Intelligence Is Stuck. Here’s How to Move It Forward” (2017)
- (This perspective is not universal, it gets into heated philosophical debates and “hard” vs. “soft” artificial intelligence, Turing Test vs. the “Chinese Room,” etc…)

What is machine learning?

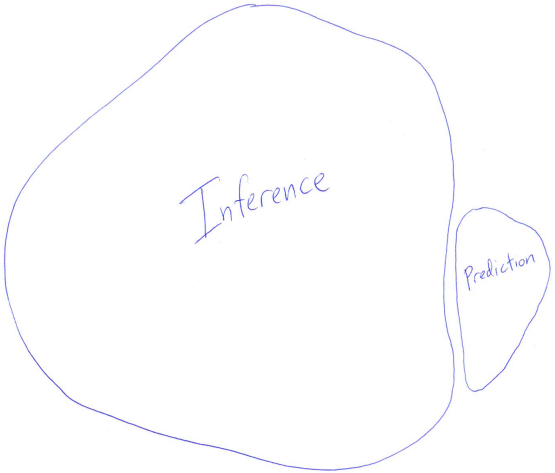
- It is quite surprising that statistical approaches, designed to uncover data-generating mechanisms with variation (and under uncertainty), could be applied to carry out operations resembling “intelligence”
- The original vision of AI had to do with modeling (and thereby reproducing) *rules and reasoning*, but it reached a dead end, at which point people found that using data and statistics could give huge gains (Halevy, Norvig, & Pereira, “The Unreasonable Effectiveness of Data”, 2009)

What is machine learning?

Machine learning relies heavily on statistical machinery, but there distinctions that are deeply important for what we do.

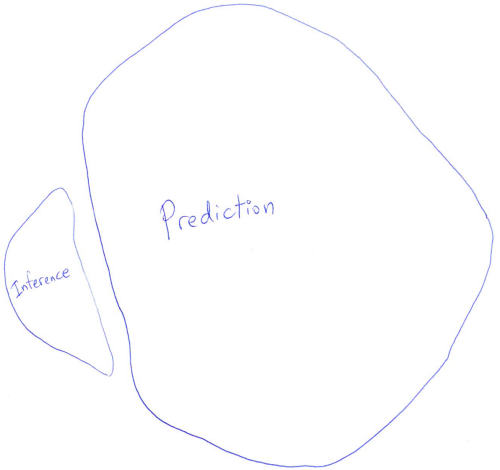
Statistics vs. machine learning

Statistics versus Machine Learning



How statisticians see the world?

Statistics versus Machine Learning

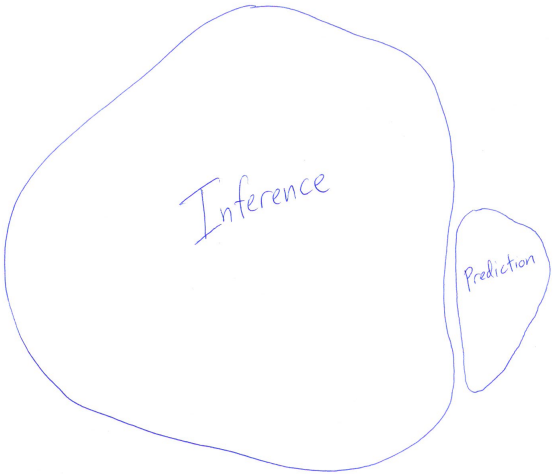


How machine learners see the world?

Diagrams: Robert Tibshirani, “Recent Advances in Post-Selection Inference” (2015)

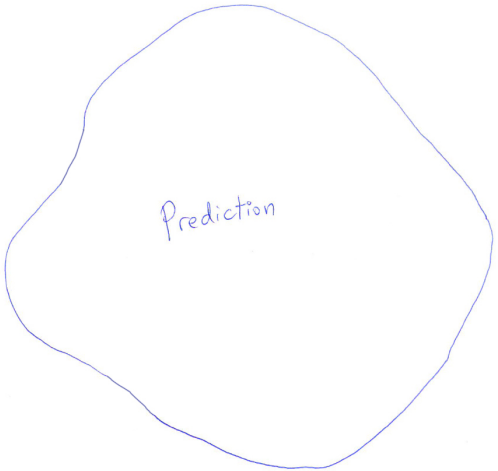
Statistics vs. machine learning

Statistics versus Machine Learning



How statisticians see the world?

Statistics versus Machine Learning



How machine learners see the world?

Diagrams: Robert Tibshirani, “Recent Advances in Post-Selection Inference” (2015)

AI Magazine Volume 18 Number 3 (1997) (© AAAI)

Does Machine Learning Really Work?

Tom M. Mitchell

■ Does machine learning really work? Yes.

…Or does it?

"performance claims can easily be taken to be an assertion about the performance of the system under general conditions. In fact, we suspect that most authors of these works had similar assumptions in mind (author's note: we did!)."

– Cohen & Ruths, "Classifying Political Orientation on Twitter: It's Not Easy!" (2013)

- Natural/mechanical systems are pretty constant in time/context
- People are not
- People also react to being modeled
- So machine learning doesn't really work (for people), and we should all go back to inference and statistics…?

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

*"There are **two cultures** in the use of statistical modeling to reach conclusions from data. One assumes that the **data are generated by a given stochastic data model**. The other uses **algorithmic models** and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to **irrelevant theory, questionable conclusions**, and has kept statisticians from working on a large range of interesting current problems."*

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

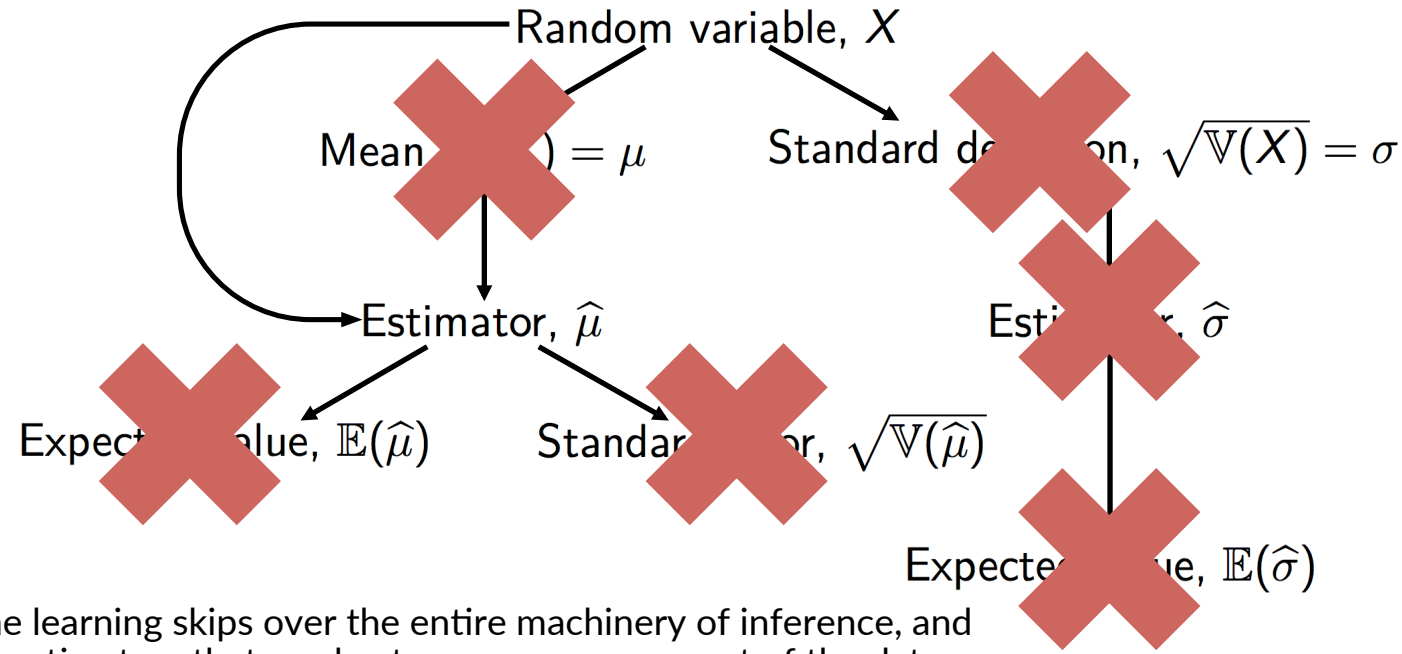
Leo Breiman

Proved prophetic in (actually) predicting the irrelevance of statistics in the face of machine learning. Makes his critiques of statistics and its focus on inference very biting.

Statistics vs. machine learning

- “the object of statistical methods is the **reduction of data**.” –R. A. Fisher, 1922
- “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P **if its performance** at tasks in T, as measured by P, **improves with experience** E.” –Tom Mitchell, 1997

Machine learning



Machine learning skips over the entire machinery of inference, and creates estimators that can best recover some aspect of the data.

To be fair, *statistical machine learning* brings much of this back in, but for the purpose of seeing what best predicts *X* (or predicts *Y*, it we are modeling the relationship between *X* and *Y*), and <u>not</u> what recovers information.

Prediction

- For inference, we estimate standard errors and do hypothesis testing as a measure of quality
- How do we judge predictions?
- The test of a prediction is how well it applies to not-yet-seen data
- We never have not-yet-seen data
- For people, not-yet-seen data may prove very different from seen data
- Data splitting through cross-validation *simulates* not-yet-seen data
- This is the sole purpose of cross-validation: use it as such

Cross-validation

- Cross validation is far from perfect.
- Data is never independent; which means we can never split data in a way that properly reflects true out-of-sample data
- *"in practice even data allocated for the sole purpose of testing is frequently reused… Such abuse of the holdout set is well known to result in significant overfitting to the holdout or cross-validation set. Clear evidence that such reuse leads to overfitting can be seen in the data analysis competitions organized by Kaggle Inc."* –Dwork et al., 2014
- In practice, if the first model we make doesn't work, we don't give up! We try others on the same test set. But this can give a *distribution over models*; according to frequentist theory, one will eventually give good test performance just by chance!

Prediction vs. Inference

- If we are doing inference (like in social science), we want to use all the data for making our estimates as good as we can! So, we don't use cross-validation
- But in practice, we don't know if our assumptions about functional forms are right
- So even for inference, it can be good to use predictive performance as a way to judge the overall fit of the model
- (If our logistic regression doesn't beat the baseline, should we really trust its inferences?)
- (This is an argument for social science using predictive performance at least for "goodness-of-fit" testing)

Correlation is not causation…

- …and, more subtly, *prediction is not explanation*
- That is, *just because we can predict well*, doesn't mean we have found *true associations*
- An easy way to understand: spurious correlations can capture underlying causal structures and may therefore lead to great predictions

When false models predict better than true ones

- Surely, though, if we have the “true” model, everything works out?
- *Not necessarily*
- *Sometimes, false models may predict better than true ones!!*
- I have a artificial example of this, but I hope this existence proof at least separates prediction and explanation in your mind
- (Unlike in physics,) “algorithmic” or “black box” models (from ML) working on superficial relationships often perform far, far better than attempts to model human behavior from first principles. This is something really weird (and disappointing).
- Joint work with Hemank Lamba (DSSG ’15)

When false models predict better than true ones

- Setup: generate standard normal X_1 through X_{12} such that X_7 is highly correlated with X_{10}, X_8 is highly correlated with X_{11}, and X_9 is highly correlated with X_{12}
- Generate y, a vector of 100 observations, as:

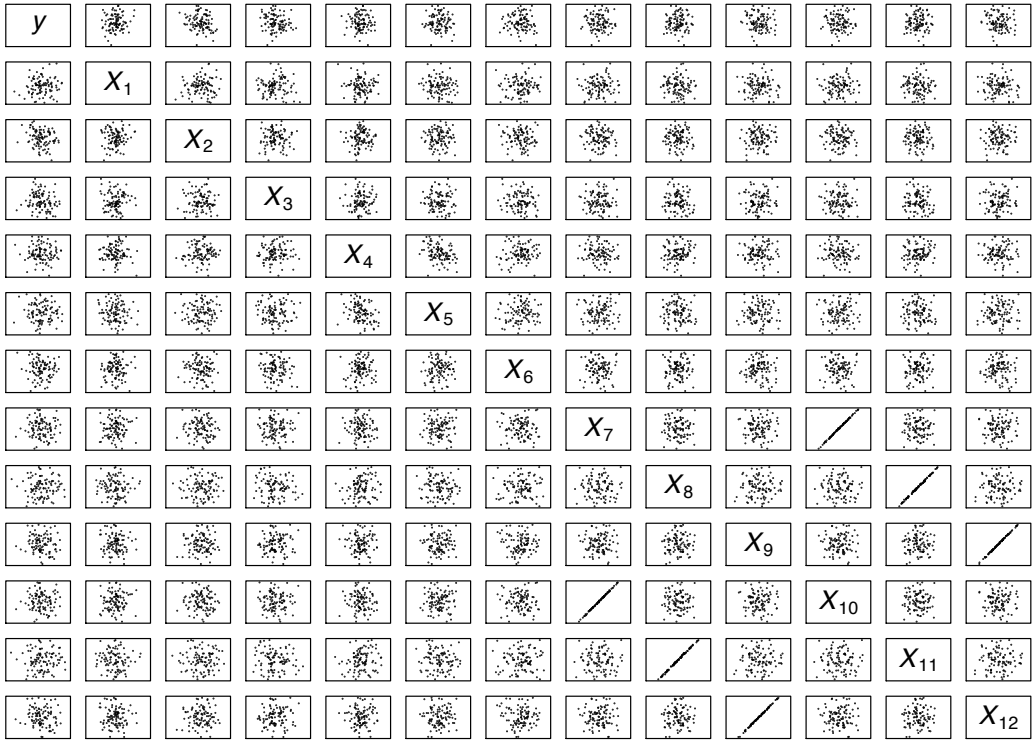
  $$y = 10^*X_1 + 10^*X_2 + \cdots + 10^*X_7 + X_8 + \cdots + X_{12} + \varepsilon$$

  where ε has a standard deviation of 4.79 (see Smueli, "To Predict or to Explain?," 2010)
- The size of the noise, ε, compared to the signal (the size of the coefficients) is the main reason why things are weird

When false models predict better than true ones

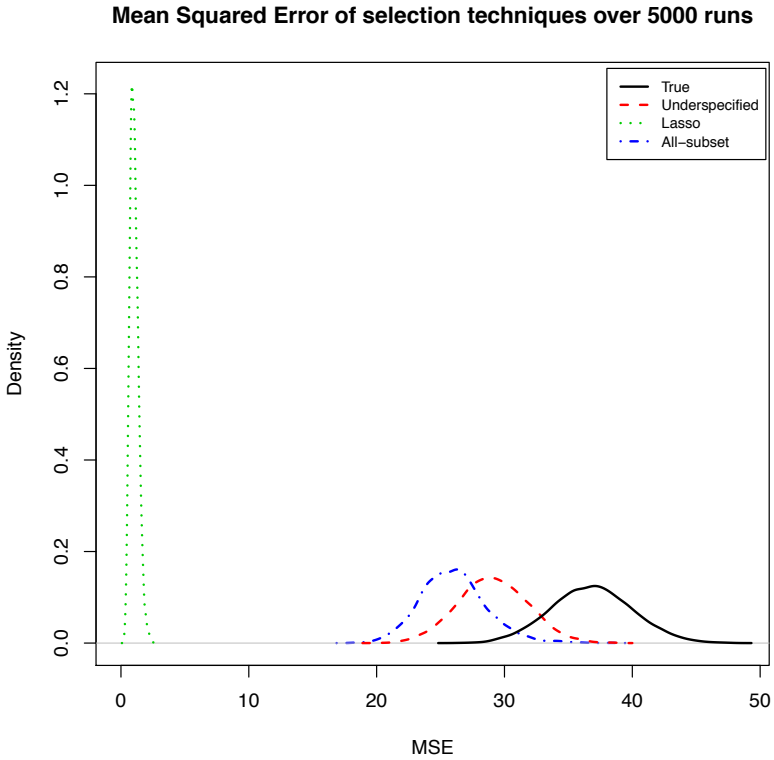
- Generate 5,000 y's (X's are fixed, only ε changes), and fit four models to each:
  1. the true model that generated the data,
  2. an "underspecified" model of only X_1 through X_9,
  3. a lasso (with bandwidth chosen through cross-validation from a validation set), and
  4. "all-subset" selection (which tries everything in the power set of all 2^{12} possible combinations of variables, optimal set again chosen through cross-validation from a validation set)
- Compute the distribution of test error of each type of model over the 5,000 draws (both mean squared error and mean absolute error)
- Does the true model do the best? No!

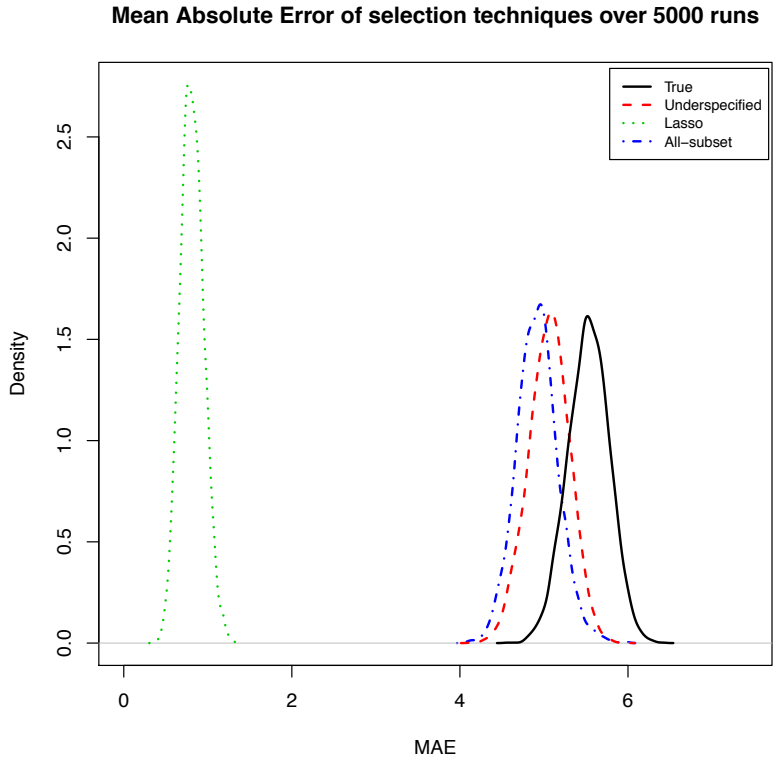
When false models predict better than true ones



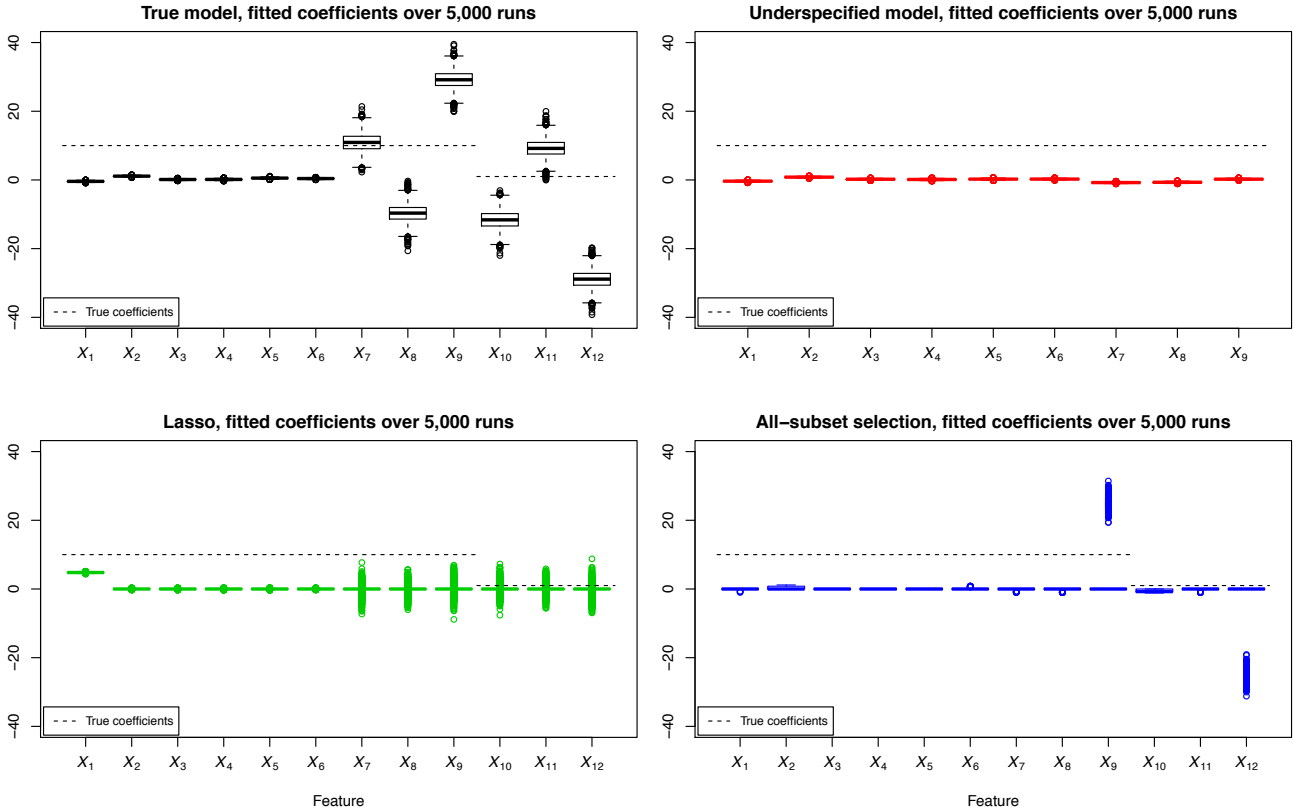
Scatter plot of *y* and *X*'s

When false models predict better than true ones





When false models predict better than true ones



Why?

- I still haven't fully figured out how to explain this at an intuitive (non-mathematical/simulation) level
- The "James-Stein estimator" showed decades ago that regularizing can sacrifice performance on individual coefficients to improve overall performance, although why this would happen is still mysterious
- *The underspecified model, all-subset selection and lasso all predict better than re-applying the model that generated the data!*
- (But nothing recovers the true coefficients)
- (Surprisingly, regularization through the lasso does better [lower test error] than even all-subset regression. Surprising because the lasso was made to be a computationally efficient approximation of all-subset selection, which is very slow)
- The *bias-variance tradeoff* is one angle: by decreasing variance and increasing bias, we do better by our measure of performance

Bias-variance tradeoff

- When we use squared error loss, it turns out we can *decompose* it into irreducible error (phenomenon), and bias squared and variance (of the estimator)
- This will make more sense next time, but for reference:

$$\begin{aligned}\mathrm{EPE}(x) &= \mathbb{E}\big[\big(Y - \widehat{f}(x)\big)^2 | X = x\big] \\ &= \mathrm{Var}(Y) + \mathbb{E}\big[\big(\widehat{f}(x) - f(x)\big)^2 | X = x\big] + \mathbb{E}\big[\big(\widehat{f}(x) - \mathbb{E}[\widehat{f}(x)]\big)^2 | X = x\big] \\ &= \sigma^2 + \mathrm{bias}^2\big(\widehat{f}(x)\big) + \mathrm{Var}\big(\widehat{f}(x)\big)\end{aligned}$$

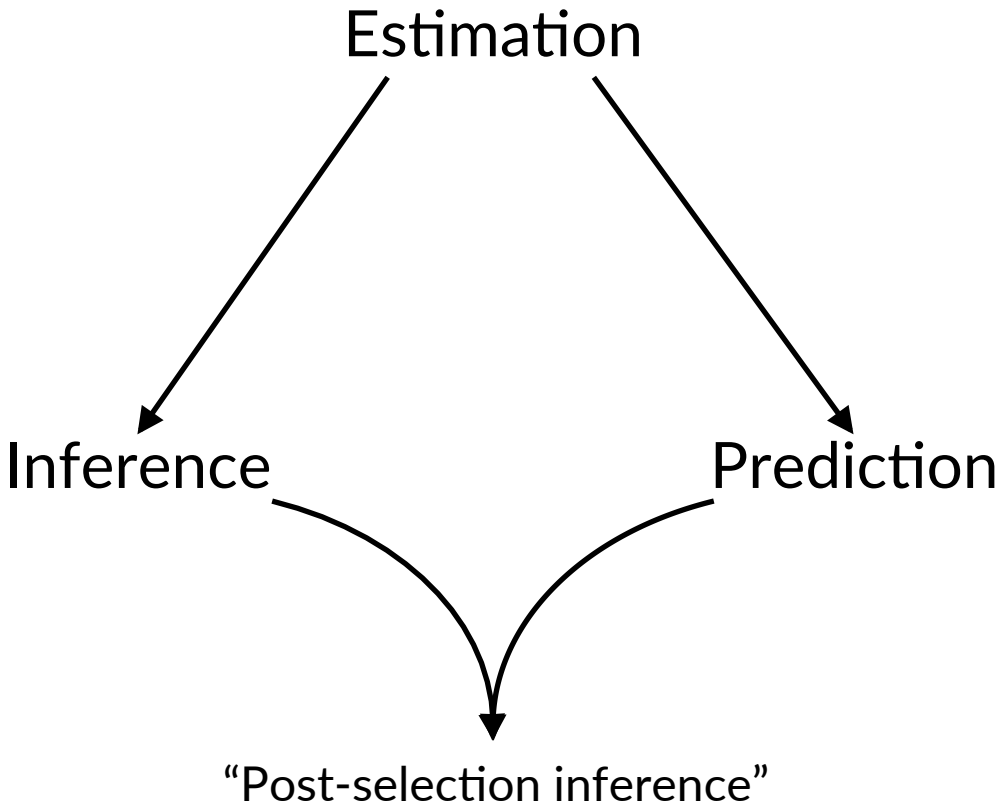
- σ^2 is the irreducible error (the variance of *Y*, beyond any signal from *X*). (Note: *X* can be a set of variables.) The bias is how far the estimator is from the true signal of *X*, and the variance is how noisy the estimator is—and this term has nothing to do with *Y*!!
- Again, note: we don't ever know the true bias. Cross validation is, in statistical terms, an *estimator* of the (true, unknown, unobservable) prediction error!
- But, be careful: the bias-variance decomposition only applies to symmetric, convex loss functions! (See Gareth M. James, "Variance and Bias for General Loss Functions," 2003)

Lessons

- Correlation is not causation, but also…
- Prediction and inference are two separate tasks
  - The variables selected out by the lasso, or random forests, as most *predictive* may not be statistically significant
  - Statistically significant variables may have little predictive power
  - Don’t conflate the two!
- False models can predict better than “true[r]” ones: prediction doesn’t mean truth!

Last note: Current work seeks to combine the two

It's incorrect to do inference after feature selection (e.g., select features, then run a model with just those features to get *p*-values), because then the inference doesn't take into account the uncertainty of the feature selection (e.g., over multiple draws of data sets from the same phenomenon, we would select out slightly different sets of features each time due to variance). But we can do some fancy math to take this into account, and widen confidence intervals to reflect the uncertainty. See Robert Tibshirani, "Recent Advances in Post-Selection Inference" (2015).



Next time:

1. Regression
   - Generalized Linear Models
   - Nonparametrics

2. Specification
   - Omitted variable bias
   - Dependent data

3. Causality
   - Graphical models

References (1/2)

Breiman, Leo. “Statistical Modeling: The Two Cultures.” *Statistical Science* 16, no. 3 (2001): 199–231. doi:10.1214/ss/1009213726

Cohen, Raviv and Derek Ruths. “Classifying Political Orientation on Twitter: It’s Not Easy!” In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 91–99. Palo Alto, California: The AAAI Press, 2014. AAAI:6128-30356-2-PB

Cox, D. R. “Role of Models in Statistical Analysis.” *Statistical Science* 5, no. 2 (May 1990): 169–174. doi:10.1214/ss/1177012165

Dwork, Cynthia, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. “Preserving Statistical Validity in Adaptive Data Analysis.” *arXiv*:1411.2664v1 (10 November 2014).

Fisher, R. A. “On the Mathematical Foundations of Theoretical Statistics.” *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222 (1922): 309–368. doi:10.1098/rsta.1922.0009

Friedman, Jerome H. “Data Mining and Statistics: What's the Connection?” In *Proceedings of the 29th Symposium on the Interface Between Computer Science and Statistics*, 1997. http://statweb.stanford.edu/~jhf/ftp/dm-stat.pdf

Gayo-Avello, Daniel. “‘I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper’: A Balanced Survey on Election Prediction using Twitter Data.” *arXiv*:1204.6441v1 (28 April 2012).

References (2/2)

Halevy, Alon, Peter Norvig, and Fernando Pereira. “The Unreasonable Effectiveness of Data.” *IEEE Intelligent Systems* (March/April 2009): 8–12. doi:10.1109/MIS.2009.36

James, Gareth M. “Variance and Bias for General Loss Functions.” *Machine Learning* 51, no. 2 (May 2003): 115–135. doi:10.1023/A:1022899518027

Marcus, Gary. “Artificial Intelligence Is Stuck. Here’s How to Move It Forward.” *New York Times* (29 July 2017). https://nyti.ms/2hav95E

Mitchell, Tom M. “Does Machine Learning Really Work?” *AI Magazine* 18, no. 3 (Fall 1997): 11–20. doi:10.1609/aimag.v18i3.1303

Mitchell, Tom M. *Machine Learning*. New York, NY: McGraw Hill, 1997.

Shalizi, Cosma. “So, You Think You Have a Power Law, Do You? Well Isn't That Special.” NY Machine Learning Meetup (18 October 2010). http://www.stat.cmu.edu/~cshalizi/2010-10-18-Meetup.pdf

Shmueli, Galit. “To Explain or to Predict?” *Statistical Science* 25, no. 3 (2010): 289–310. doi:10.1214/10-STS330

Tibshirani, Robert. “Recent Advances in Post-Selection Inference.” Breiman Lecture, NIPS 2015 (9 December 2015) http://statweb.stanford.edu/~tibs/ftp/nips2015.pdf

Wasserman, Larry A. “Rise of the Machines.” In *Past, Present, and Future of Statistical Science*, 525–536. Boca Raton, FL: Chapman and Hall/CRC, 2013. http://www.stat.cmu.edu/~larry/Wasserman.pdf